Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-03-22 14:58:39
REFERENCE NO: 188

Author Names & Affiliations

- Andrew Medford - Georgia Institute of Technology

Contact Email Address (for NSF use only)

(Hidden)

Research Domain, discipline, and sub-discipline

Chemical Engineering, Computational Catalysis

Title of Submission

Uncertainty quantification, analysis, and propagation in computational chemistry.

Abstract (maximum ~200 words).

Computational chemistry is ubiquitous, with tens of thousands of papers published annually and wide-ranging applications in energy materials, biological sciences, and nuclear technology. There has been significant focus on the development of methods and software packages for more accurate and faster calculations of chemical properties, but relatively little attention has been paid to systematic quantification of uncertainty. This is particularly problematic since density functional theory (DFT), the workhorse method in computational chemistry, is not systematically improvable. Due to this, many calculations are of questionable and unquantified accuracy. Recent work has identified various strategies for systematically quantifying uncertainty in fitted DFT functionals and molecular dynamics force-fields, but these techniques often require significant modification to the canonical implementations and are hence not integrated into existing cyberinfrastructure. Furthermore, even when uncertainty is quantified there is a significant lack of tools for analyzing the uncertainty and its impacts on scientific conclusions. The development and implementation of novel tools for visualization, analysis, and propagation of uncertainty in fitted density functionals, molecular dynamics force-fields, and multi-scale chemical models will facilitate the development of more accurate techniques and lead to more robust conclusions from computational chemistry data.

Question 1 Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

The quantification of uncertainty is a relatively new area in the field of molecular sciences, however there have been several works that demonstrate the potential for the quantum-mechanical technique of DFT, the classical force-field based approach, and the advantages of propagating uncertainty in multi-scale models.

In the case of DFT, the exchange-correlation approximation is known to be inaccurate, but its accuracy is difficult to assess in a general

sense. For example, more physically complex approximations have larger errors than simpler ones for certain systems [1]. This was addressed by Jacobsen, Sethna, Norskov, and Bligaard in a series of papers where a Bayesian parameter estimation scheme was devised to map errors from the output energy space to the underlying parameter space of the functional [2-4]. In this way it is possible to generalize the errors between a fitted functional and a known set of highly accurate energies to other systems where the functional was not trained. This approach leads to an ensemble of DFT energies for any given system, equivalent to a systematic sensitivity analysis of energy with respect to functional parameters. Similar techniques have been adopted by others in order to estimate errors for other classes of systems [5]. While effective, these approaches are limited by i) the flexibility of the underlying model and ii) the subset of training energies used. Currently, design of such functionals is highly centralized and time consuming due to the fact that there is no open-source framework for functional fitting and design. Furthermore, the resulting models are often opaque, making it difficult for researchers in the field to interpret the source of uncertainty for a given energy. This hinders the adoption of these approaches despite their strong ability to improve the robustness of DFT calculations.

Another emerging area of interest is data-driven classical force-fields derived from large amounts of DFT data [6-8]. These force-fields are much more flexible than "physically derived" force-fields, and much cheaper than DFT calculations, enabling the investigation of much larger time and length scales for complex systems like surfaces and reactions where bond formation occurs in heterogeneous environments. Uncertainty quantification is critical for these machine-learning force fields to ensure that the test systems do not move far beyond the domain of the training system, effectively enabling researchers to know when a molecular dynamics simulation moves out of the scope of the training data. However, the most popular neural network forcefields typically do not contain any measure of uncertainty [6], while the force-fields that do account for uncertainty are less prevalent or limited to systems with a single atomic species [7-8]. Furthermore, the ability to conduct molecular dynamics simulations with uncertain energies/forces has only recently emerged [9], and is not implemented in any openly available molecular dynamics software packages and is hence only rarely utilized.

An additional frontier for uncertainty analysis in the chemical sciences is propagation through multi-scale models of chemical reaction kinetics [10-12]. Although the number of studies is limited, initial results suggest that error propagation through kinetic models is non-trivial, and that significant cancellation of error occurs due to correlations in the underlying parameters [10-11]. Furthermore, interesting approaches have been developed to utilize probability distributions from both computational and experimental techniques in order to assess uncertainty and/or model reliability [11-12]. This quantitative coupling of theory and experiment holds great promise for creating more impactful models, but relies on significant mathematical analysis that is not approachable for the average researcher in the chemical sciences. Given the limited number of examples, the generality of the effects of error cancellation are unknown, and agreement between experiment and theory is assessed only qualitatively in nearly all cases.

[1] M. G. Medvedev, I. S. Bushmarinov, J. Sun, J. P. Perdew, and K. A. Lyssenko, “Density functional theory is straying from the path toward the exact functional,” Science, vol. 355, no. 6320, pp. 49–52, Jan. 2017.

[2] S. L. Frederiksen, K. W. Jacobsen, K. S. Brown, and J. P. Sethna, “Bayesian Ensemble Approach to Error Estimation of Interatomic Potentials,” Physical Review Letters, vol. 93, no. 16, Oct. 2004.

[3] J. J. Mortensen, K. Kaasbjerg, S. L. Frederiksen, J. K. Nørskov, J. P. Sethna, and K. W. Jacobsen, “Bayesian Error Estimation in Density-Functional Theory,” Physical Review Letters, vol. 95, no. 21, Nov. 2005.

[4] J. Wellendorff, K. T. Lundgaard, A. Møgelhøj, V. Petzold, D. D. Landis, J. K. Nørskov, T. Bligaard, and K. W. Jacobsen, “Density functionals for surface science: Exchange-correlation model development with Bayesian error estimation,” Physical Review B, vol. 85, no. 23, Jun. 2012.

[5] G. N. Simm and M. Reiher, “Systematic Error Estimation for Chemical Reaction Energies,” Journal of Chemical Theory and Computation, vol. 12, no. 6, pp. 2762–2773, Jun. 2016.

[6] J. Behler, S. Lorenz, and K. Reuter, “Representing molecule-surface interactions with symmetry-adapted neural networks,” The Journal of Chemical Physics, vol. 127, no. 1, p. 014705, Jul. 2007.

[7] V. Botu, R. Batra, J. Chapman, and R. Ramprasad, “Machine Learning Force Fields: Construction, Validation, and Outlook,” The Journal of Physical Chemistry C, vol. 121, no. 1, pp. 511–522, Jan. 2017.

[8] A. P. Bartók and G. Csányi, “Gaussian approximation potentials: A brief tutorial introduction,” International Journal of Quantum

Chemistry, vol. 115, no. 16, pp. 1051–1057, Apr. 2015.

[9] A. V. Tran and Y. Wang, “Reliable Molecular Dynamics: Uncertainty quantification using interval analysis in molecular dynamics simulation,” Computational Materials Science, vol. 127, pp. 141–160, Feb. 2017.

[10] A. J. Medford, J. Wellendorff, A. Vojvodic, F. Studt, F. Abild-Pedersen, K. W. Jacobsen, T. Bligaard, and J. K. Norskov, “Assessing the reliability of calculated catalytic ammonia synthesis rates,” Science, vol. 345, no. 6193, pp. 197–200, Jul. 2014.

[11] J. E. Sutton, W. Guo, M. A. Katsoulakis, and D. G. Vlachos, “Effects of correlated parameters and uncertainty in electronic-structure-based chemical kinetic modelling,” Nature Chemistry, vol. 8, no. 4, pp. 331–337, Feb. 2016.

[12] E. Walker, S. C. Ammal, G. A. Terejanu, and A. Heyden, “Uncertainty Quantification Framework Applied to the Water–Gas Shift Reaction over Pt-Based Catalysts,” The Journal of Physical Chemistry C, vol. 120, no. 19, pp. 10328–10339, May 2016.

Question 2 Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

The largest gap in cyberinfrastructure for dealing with uncertainty in the chemical sciences is the lack of openly available and user-friendly software packages. Properly accounting for uncertainty in DFT, molecular dynamics, and multi-scale models requires significant statistical and mathematical knowledge that is far beyond the expertise of most researchers in the field. Most groups working in this area utilize in-house code, but this is only rarely made available, and even less commonly documented. Developing software that is user-friendly enough to be utilized by other groups will greatly accelerate the adoption of these techniques, in a similar way to the explosion of computational chemistry that accompanied the release of Gaussian and VASP software packages for quantum chemistry and DFT. Given the significant amount of uncertainty that exists in DFT, molecular dynamics force fields, and multi-scale reaction models it is critical that more researchers learn how to quantify and understand this uncertainty in order to draw more robust and reliable scientific conclusions from their calculations. Some examples of software packages that would be particularly impactful are:

1) Flexible frameworks for fitting DFT exchange-correlation functionals with uncertain parameters.

2) Implementation of efficient tools for training machine-learning molecular dynamics force fields with error estimates.

3) Development of molecular dynamics packages that account for uncertainty in forces and energies.

4) Kinetic simulation software that propagates correlated distributions of rate/equilibrium constants to distributions of reaction rates (or vice versa).

In addition to the technical components of these software packages considerable attention should be paid to their documentation and interface. Understanding uncertain quantities is often not intuitive, and the development of advanced visualization techniques and interactive tutorials will significantly lower the adoption barrier for researchers in the field with less statistical knowledge. The expertise of computer scientists working in human-computer interaction and data visualization is expected to be particularly valuable in this regard.

Question 3 Other considerations (maximum ~1200 words, optional): Any other relevant aspects, such as organization, process, learning and workforce development, access, and sustainability, that need to be addressed; or any other issues that NSF should consider.

Emphasis should be placed on the development of open-source tools that will be released to the community. Collaborations between researchers in chemical and computational sciences will help ensure the development of high-quality software that is useful to the

community. Involvement of the chemical industry will also help ensure that the tools are adopted outside of academia, will help re-train the current workforce to use novel data science tools, and will ensure that the software that is developed is sufficiently user-friendly to be adopted by the wider community.

Consent Statement

- "I hereby agree to give the National Science Foundation (NSF) the right to use this information for the purposes stated above and to display it on a publically available website, consistent with the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode)."